PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : **G01N 27/447**	**A1**

(11) International Publication Number: **WO 98/00708**

(43) International Publication Date: 8 January 1998 (08.01.98)

(21) International Application Number: PCT/CA97/00463

(22) International Filing Date: 26 June 1997 (26.06.97)

(30) Priority Data:
08/670,534   27 June 1996 (27.06.96)   US

(60) Parent Application or Grant
  (63) Related by Continuation
    US
    Filed on   08/670,534 (CIP)
       27 June 1996 (27.06.96)

(71) Applicant (for all designated States except US): VISIBLE GENETICS INC. [CA/CA]; 700 Bay Street, Box 333, Toronto, Ontario M5G 1Z6 (CA).

(72) Inventors; and
(75) Inventors/Applicants (for US only): GILCHRIST, Rodney, D. [CA/CA]; 1114 Edgehill Place, Oakville, Ontario L6M 2E7 (CA). CHI, Vrijmoed [CA/CA]; Unit 54, 3125 Fifth Line West, Mississauga, Ontario L5L 3S8 (CA).

(74) Agent: DEETH WILLIAMS WALL; National Bank Building, Suite 400, 150 York Street, Toronto, Ontario M5H 3S5 (CA).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).
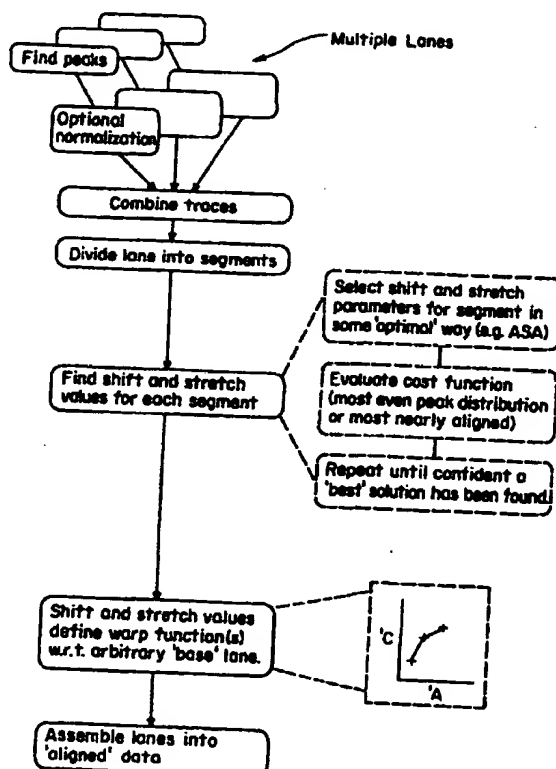
Published
With international search report.

(5,916,747)

(54) Title: METHOD AND APPARATUS FOR ALIGNMENT OF SIGNALS FOR USE IN DNA BASE-CALLING

(57) Abstract

Data traces from four channels of an automated electrophoresis detection apparatus are aligned by identifying peaks in each of the four data traces; optionally normalizing the data traces to achieve a uniform peak height; combining the four data traces in an initial alignment; and determining coefficients of shift and stretch for selected data points within each data trace. The coefficients are determined by optimizing a cost function which reflects the extent of overlap of peaks in the combined normalized data traces to which the coefficients have been applied. The cost function is optimized when the extent of overlap is at a minimum. The coefficients are then used to generate a warp function for each data trace. These warp functions are applied to their respective data traces to produce four warped data traces which are aligned to form an aligned data set. The aligned data set may be displayed on a video screen of a sequencing apparatus, or may be used as the data set for a base-calling process.

METHOD AND APPARATUS FOR ALIGNMENT OF SIGNALS
FOR USE IN DNA BASE-CALLING

DESCRIPTION

BACKGROUND TO THE INVENTION

This invention relates to a method of processing output signals from an automated electrophoresis detection apparatus, and to an apparatus which employs this method for sequencing nucleic acids.

5   One of the steps in nucleotide sequence determination of a subject nucleic acid molecule is interpretation of the pattern of nucleic acid fragments which results from electrophoretic separation of fragments, or reaction products, of a DNA sequencing reaction (the "fragment pattern"). The interpretation, colloquially known as "base calling", involves determination from the recorded fragment pattern of the order of four nucleotide bases, A

10   (adenine), C (cytosine), G (guanine) and T (thymine) for DNA or U (uracil) for RNA in the subject nucleic acid molecule.

The chemistry employed for a DNA sequencing reaction using the dideoxy (or chain-termination) sequencing technique is well known, and was first reported by Sanger et al. (Proc. Natl. Acad. Sci. USA 74: 5463-5467 (1977)). Four samples of nucleic acid fragments

15   (terminating in A, C, G, or T(U) respectively in the Sanger et al. method) are loaded at a loading site at one end of an electrophoresis gel. An electric field is applied across the gel, causing the fragments to migrate from the loading site towards the opposite end of the gel. During this electrophoresis, the gel acts as a separation matrix. The fragments, which in each sample are of an extended series of discrete sizes, separate into bands of discrete species in a

20   lane along the length of the gel. Shorter fragments generally move more quickly than larger fragments.

If the DNA fragments are labeled with a fluorescent label, an automated electrophoresis detection apparatus (also called a "DNA sequencer") can be used to detect the passage of migrating bands in real time. Existing automated DNA sequencers are available from Applied

25   Biosystems, Inc. (Foster City, CA), Pharmacia Biotech, Inc. (Piscataway, NJ), Li-Cor, Inc. (Lincoln, NE), Molecular Dynamics, Inc. (Sunnyvale, CA) and Visible Genetics Inc. (Toronto). Other methods of detection, based on detection of features inherent to the subject

molecule, such as detection of light polarization as disclosed in US Patent No. 5,543,018 which is incorporated herein by reference, are also possible.

A significant problem in determining a DNA sequence, encountered particularly with high speed DNA sequencing and in sequencing apparatus which do not combine the four sets of sequencing reaction products in a single lane, is alignment of data signals from the four different output channels of an automated DNA sequencing apparatus. Once data is aligned properly, it is relatively straight-forward to base-call it. However, this initial step can be very challenging since the output signal may be erratically shifted and/or stretched as a result of chemistry and gel anomalies. A reliable method of aligning data, that can produce data which takes into account non-linear shifting and stretching of signal output, is highly desirable particularly for high-speed DNA sequencing.

Existing prior art determinants in this field are very limited. Existing automated sequencers traditionally operate at voltages low enough that non-linear shifting is avoided. The use of low voltages, however, limits the speed with which separation of sequencing fragments into discrete bands can be accomplished.

Published methods of computer assisted base calling include the methods disclosed by Tibbetts and Bowling (US Pat. No. 5,365,455) and Dam et al (US Pat. No. 5,119,316) which patents are incorporated herein by reference. Both patents assume alignment of output signals and address only aspects of base-calling from the aligned signals.

It is an object of the present invention to provide a method of aligning real-time signals from the output channels of an automated electrophoresis apparatus.

It is a further object of the invention to provide an improved method of base-calling an DNA signal sequence aligned according to the invention.

It is still a further object of the invention to provide an apparatus for sequencing nucleic acids which utilizes the improved method in accordance with the invention for aligning real-time signals from the output channels of an automated electrophoresis apparatus.

SUMMARY OF THE INVENTION

These and other objects of the invention are achieved using a method for aligning data traces from four channels of an automated electrophoresis detection apparatus, each channel detecting the products of one of four chain-termination DNA sequencing reactions, whereby

said four channels together provide information concerning the sequence of all four bases within a nucleic acid polymer being analyzed, comprising the steps of:

(a)   identifying peaks in each of the four data traces;

(b)   normalizing the height of said peaks in each of said data traces to a common value to generate four normalized data traces if the peaks are not of substantially equal height;

(c)   combining the four normalized data traces in an initial alignment;

(d)   determining coefficients of shift and stretch for selected data points within each normalized data trace, said coefficients optimizing a cost function which reflects the extent of overlap of peaks in the combined normalized data traces to which the coefficients have been applied, said cost function being optimized when the extent of overlap is at a minimum;

(e)   generating warp functions for the normalized data traces from the coefficients of shift and stretch determined for the selected data points;

(f)   applying the warp functions to the respective data trace or normalized data trace to produce four warped data traces; and

(g)   assembling the four warped data traces to form an aligned data set.

The aligned data set may be displayed on a video screen of a sequencing apparatus, or may be used as the data set for a base-calling process.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows the alignment process of the invention in flow chart form;

Fig. 2 illustrates a preliminary (unaligned) signal recorded from an automated DNA sequencing apparatus over 15 seconds;

Fig. 3 illustrates the detection of peaks from a preliminary signal;

Fig. 4 illustrates a normalized data stream;

Figs. 5A and 5B illustrate a normalization method employed for peaks and valleys;

Figs. 6A and 6B illustrate the preliminary alignment of four normalized data streams representing time based alignment of each of the A, C, G and T output channels;

Fig. 7 illustrates alignment of the 4 normalized data streams with respect to minimization of the "cost" function;

Figs. 8A and 8B shows exemplary warp functions for data points in the C, G and T traces relative to the A trace;

Fig. 9 illustrates a method for determining a standard gaussian peak width;

Fig. 10 shows an example of the classification of singletons using the features of peaks in the detected signal;

Fig. 11 illustrates a base-calling method useful with the aligned data traces; and

Fig. 12 shows an apparatus in accordance with the invention.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a method of aligning data traces which can be used to align data traces from an automated electrophoresis detection apparatus for use in base-calling. In accordance with the invention, an optimization routine such as a simulated annealing algorithm is used to determine coefficients of stretch and shift to align data traces, each representing one of the four chain-termination DNA sequencing reactions. With high quality data trace alignment, base-calling may proceed with a high degree of accuracy and repeatability.

"Data trace" as used in the specification and claims of this application refers to the series of peaks and valleys representing the migrating bands of oligonucleotide fragments produced in one chain termination sequencing reaction and detected in a DNA sequencer. Such data traces are sometimes referred to as sequence chromatograms or a chromatographic trace. The data trace may be either a raw data trace or a "conditioned" data trace.

"Shift" as used in the specification and claims of this application is interchangeable with "offset" and refers to the number of data points which the signal output is displaced, either positively or negatively from its original position.

"Stretch" as used in the specification and claims of this application refers to increase or decrease in spacing between data points of a signal output relative to the original spacing. The stretch may be constant across a segment of data or may follow a second or higher order polynomial.

A "warp function" as used in the specification and claims of this application is an instantaneous representation of optimized shift and stretch at each data point. The warp function can be represented graphically either as a data point-for-data point plot for one data

- 5 -

trace versus a standard or second data trace, or as a plot of displacement as a function of data point number. Standard traces may be a trace representing an average of multiple experimental runs, a base separation function, or a standard derived from a text sequence. The warp function is applied to a data trace to obtain a "warped data trace."

5   Fig. 1 shows the alignment process of the invention in flow chart form. As shown, the first step in the method is the identification of peaks in each of several (generally four) data traces. In some cases, these traces are next normalized to a constant height. This step may be unnecessary if the DNA Sequencer and chemistry used produce substantially uniform peak sizes (homozygous peak heights within about 10 %) or if the traces are nearly aligned from the

10   sequencer. The term "normalized data traces" refers to data traces having this level of consistency, whether as a result of normalization step or as a result of sufficient initial uniformity. Normalization may also be unnecessary, even for non-uniform peak sizes, if the cost function used in the alignment is independent of peak height and area.

The data traces are then combined in an initial alignment and divided into a plurality of

15   segments or windows. For each segment, optimized stretch and shift values are determined using an iterative optimization routine which converges towards a "best" solution for alignment of the data traces. The quality of the various trial alignment is assessed for comparison purposes using a cost function. Various cost functions can be employed including (1) a cost function which minimizes overlap of the data traces combined within the segment;

20   and (2) a cost function which maximizes the regularity of peak distribution the segment.

Once a "best" solution has bene found, the shift and stretch values are used to define warp functions. These warp functions are applied to the individual data traces to produce warped data traces which are assembled as aligned data useful for base-calling.

To illustrate the use of this method, Fig. 2 shows a typical raw data trace for one of

25   the four chain-termination DNA sequencing reactions, as detected on a Visible Genetics Inc. MicroGene Blaster™. The X-axis represents time while the Y axis represents fluorescence detection. The data trace reveals a series of bands of fluorescent molecules passing through the detection site, as expected from a typical chain-termination DNA sequencing reaction.

Fig. 2 shows several of the features which complicate the use of raw data for base-

30   calling, and illustrates the need for the present invention to provide appropriately aligned signals for this purpose. In particular, as reflected in Fig. 2, the fluorescence intensity can vary

- 6 -

from one band to another. In addition, not all bands are fully resolved, and the spacing between adjacent bands is not always an integral multiple of the theoretical spacing between adjacent bands. The present invention provides a method for converting this raw signal, and its three counterpart signals for the other three sequencing reactions, into aligned data which is highly suitable for base-calling.

The data trace which is processed in accordance with the method of the invention is preferably a signal collected using the fluorescence detection apparatus of an automated DNA sequencer. However, the present invention is applicable to any data set which reflects the separation of oligonucleotide fragments in space or time, including real-time fragment patterns using any type of detector, for example a polarization detector as described in US Patent No. 5,543,018; densitometer traces of autoradiographs or stained gels; traces from laser-scanned gels containing fluorescently-tagged oligonucleotides; and fragment patterns from samples separated by mass spectrometry.

In the method of the invention, four data traces, one for each sequencing reaction, are normalized as needed to correct for variations in peak height using the procedure described below. The normalized data traces are then used to determine a series of stretch and shift coefficients, which are then applied to the normalized data traces to arrive at aligned data traces.

Prior to normalizing and aligning the data traces for the four sequencing reactions using the method of the invention, however, it may be advantageous to condition the signal, although this step is not required. This conditioning can be done, for example, using conventional baseline correction and noise reduction techniques to yield a "conditioned" data trace. As is known in the art, three methods of signal processing commonly used are background subtraction, low frequency filtration and high frequency filtration, and any of these may be used, singly or in combination to produce a conditioned signal to be used as a conditioned data trace in the method of the invention.

Preferably, the data is conditioned by background subtraction using a non-linear filter such as an erosion filter, with or without a low-pass filter to eliminate systemic noise. The preferred low-pass filtration technique is non-causal gaussian convolution.

After any needed conditioning of the data trace is performed, the data trace is normalized as needed to generate a "normalized data trace" which is used to determine

- 7 -

coefficients of stretch and shift for base-calling. The normalization process includes the following steps.

Firstly, the data trace (raw or conditioned) is searched for peaks. Peaks can be identified as the middle data point of three consecutive data points wherein the inside data
5   point is higher than the two outside data points. Fig. 3. More sophisticated methods of peak detection are also possible. For example, a preferred method involves using the "three-point" method to segment the data trace, and then joining the segments. A trace feature is assigned as an actual peak whenever the difference between a maximum and an adjacent minimum exceeds a threshold value, e.g., 5%. A minimum peak height from the base-line may also be
10   required to eliminate spurious peaks.

An exception is made for the so-called "primer peak" and "termination peak" which are found in some variations of the chain-termination sequencing method. These peaks comprise a large volume of unreacted primer, which tends to interfere with base-calling around the shorter chain-extension products, and a large volume of the complete sequence which may
15   interfere with base-calling around the longest chain-extension products. These peaks are identified and eliminated from consideration either on the basis of their size, their location relative to the start and end of the electrophoresis process, or some other method.

After elimination of the primer and termination peaks, the data trace is normalized so that all of the identified peaks have the same height which is assigned a common value, e.g., 1.
20   (Fig. 4). This process reduces signal variations due to chemistry and enzyme function, and works effectively for homozygous samples and for many heterozygotes having moderate, i.e., less than about 5 to 10%, heterozygosity in a 200 base pair or larger region being sequenced. It will be appreciated that when the data trace has substantially uniform peak height, normalization can be omitted.

25   To normalize the data trace, the points between each peak are assigned a numerical height value based on their position in the data trace relative to a hypothetical line joining consecutive peaks and the base line of the signal. Thus, as shown in Fig. 5A, the valley between peaks 1 and 2 has a minimum at a point which is approximately 25% of the distance from the baseline B to the line C joining peaks 1 and 2. The minimum of this valley is
30   therefore assigned a value of about .25. (See Fig. 5B). Similarly, the valley between peaks 2 and 3 has a minimum at a point which is approximately 80% of the distance from the baseline

- 8 -

B to the line D joining peaks 2 and 3. The minimum of this valley is therefore assigned a value of about 0.8 in the normalized data trace.

The next step in the method of the invention is the determination of shift and stretch coefficients for a set of four normalized data traces, one for each sequencing reaction. This is accomplished by combining the four normalized data traces as shown diagrammatically in Figs. 6A and 6B, and determining coefficients of shift and stretch for selected data points within each normalized data trace which optimize a "cost" function. The cost function generally reflects the suitability of a trial alignment resulting from the application of the trial stretch and shift coefficients to the combined data traces. One type of cost function evaluates the extent of overlap of peaks in the combined normalized data traces to which the coefficients have been applied, and is considered optimized when the extent of overlap is at a minimum. It will be understood that the terms "best solution," "optimized" and "minimum" as used herein do not require absolute optimization to an absolute minimum, which could require unreasonably long periods of analysis time, but only require a practical level of optimization sufficient to achieve satisfactory alignment of the data traces for base-calling.

In one embodiment, the cost function measures the total area above the combined normalized data traces, i.e., the dotted area in Fig. 5B. The normalized data traces are then shifted and stretched in an effort to minimize the value of "cost." In a second embodiment, "cost" is set equal to the area below the combined normalized data traces, and the data traces are then shifted and stretched to maximize the value of cost (i.e. reducing the overlap of the peaks). However, it has been found empirically that this latter approach emphasizes less valuable features of the data traces than using the area above the curves as the cost function.

In another embodiment, the cost function combines the total area above the combined data traces (the dotted regions in Fig. 6B) with the area below the highest edge formed by the combined data traces and above the second highest edge of the combined data traces (the lined regions in Fig. 6B). The cost function is at a minimum when the first of these areas is minimized and the second of these areas is maximized. When the coefficients which produce the minimum cost are applied to the normalized data traces, an aligned normalized data set as shown in Fig. 7 results.

Where the likely sequence of the DNA being analyzed is known, for example in repetitive diagnostic applications, the cost function used in the method of the invention may also compare the experimental data traces to a set of model data traces. In this case, a suitable cost function is the area between the experimental data trace and a corresponding data trace from a set of model data traces that have been prealigned. The cost function is optimized when the area between the traces is at a minimum. Standard sets of model data traces comprised of random distributions of bases might also be used in such a comparison, although this increases the cost function space.

Another form of cost function that can be used in the method of the invention evaluates the regularity of peak spacing achieved by the trial alignment. Such cost functions do not require normalization, because they do not depend on peak height or area. An example of such a function would be a function determining the standard deviation of the distance between adjacent peaks, excluding obvious outliers, with a low value for the cost function being desirable. Another example of a cost function of this type determines the distance between each peak and the closest base position within a postulated regular array of peak spacings. The cost function is optimized when the standard deviation of these distances for all of the peaks is at a minimum.

The postulated peak spacing in this array is tested over a several trial values established for a specific instrumental and experimental configuration, thus adding an additional dimension to the cost function space. For example, where it is found that the normal peak spacing is 7 data points, trial peak spacings of 3 to 9 data points are suitably tested for each trial alignment to determine the lowest value of the cost function.

It will further be appreciated that the various cost functions discussed above can be used individually or in combination since they all work towards a common goal, the optimum alignment of the data traces. Thus, alternative cost functions can be created as the product or dividend that results when two or more of the cost functions are combined.

Because the optimization of the stretch and shift coefficients is an iterative process involving the testing of many combination of trial values, when the cost space is too large reasonably to permit full-sampling of the cost space, it is desirable to use an optimization routine which facilitates convergence towards an acceptable set of coefficients. It will be appreciated by persons skilled in the art that there are many types of optimization routines

using random or directed sparse sampling techniques which might be employed, including genetic algorithms and Monte Carlo techniques. A preferred method for determining the coefficients of stretch and shift that yield the optimum value of "cost" employs "simulated annealing."

Simulated annealing is a mathematical method of searching a broad parameter space for the "best fit" result, without having to test every member of the parameter space. This method is particularly relevant to signal alignment problems in high speed DNA sequencers because the parameter space for possible coefficient of stretch and shift needs to be extremely broad to accommodate the unpredictability and variation within each electrophoresis run.

The preferred manner of performing a simulated annealing calculation employs a computer. Many computer algorithms employing simulated annealing are known and available to those skilled in the art. Of particular interest are papers by Ingber et al.:

Ingber, A.L., "Very fast simulated re-annealing," *J Mathl. Comput. Modelling* 12(8): 967-973 (1989);

Ingber, A.L. et al., "Genetic algorithms and very fast simulated reannealing: A comparison," *J Mathl. Comput. Modelling* 16 (11): 87-100 (1992);

Ingber, A.L., "Simulated annealing: Practice versus theory," *J Mathl. Comput. Modelling* 18(11): 29-57 (1993); and

Ingber, A.L., "Adaptive simulated annealing (ASA): Lessons learned, " *J Control and Cybernetics* 25(1): 33-54 (1996).

Each of these papers is incorporated herein by reference.

As employed in the present invention, simulated annealing determines coefficients of stretch and shift for signal outputs from a Visible Genetics MicroGene Blaster™ as follows. Consider a window of normalized signal output. The normalized data trace consists of a series of data points generated every 0.5 seconds. Each peak consists of about 6 to 7 data points, and requires 3 to 3.5 seconds to pass through the detection zone. For convenience of illustration, the window shown represents about 90 data points or about 15 peaks. In a preferred embodiment, however, windows of 180 to 350 data points, representing about 30 to 50 peaks, most preferably of about 250 data points representing about 40 peaks are used.

One window is created for data from each of the four data traces, and the windows would be initially aligned on the basis of data-point number. Superimposing the four windows

reveals a non-minimized "cost" result, that is, the amount of "cost" area is greater than it could be.

It is found that a first order equation can be applied to each point of a data trace within the window to modify its position and change the cost area:

$$Y = mX + b$$

where Y = the new position of point X, b = offset (or shift), and m = stretch. A second order or higher equation (with coefficients in addition to stretch and shift being determined) is apparently not necessary to obtain satisfactory results, although second or higher order equations may be used for more sophisticated analyses.

The parameter space for b and m is empirically selected. The offset coefficient (b) usually falls within 30 data points of the initial time-based alignment. Offset steps of 0.35 data points are suitably employed, thus providing a parameter space of 200 choices. For MicroGene Blaster™ data, this offset represents about 5 data peaks. The stretch coefficient (m) usually falls within 5%. These steps are suitably made in 0.66% amounts, thus requiring 15 steps to cover the whole range. The total range of parameters for stretch and shift for each signal output is therefore 3000 (200 * 15). The range of parameters for alignment of three channels with respect to the fourth channel is therefore 27 billion (3000^3).

For each combination of the six parameters tested, the coefficient are applied to the three signal output functions and the functions are stretched and/or shifted accordingly. The adjusted functions are superimposed with the fourth signal output and the "cost" area is re-calculated. The cost will either be lower, higher or the same as before.

Efficient selection of the six parameters is crucial for the discovery of the parameters which provide the lowest cost. The simulated annealing theory selects parameters for testing according to a variation of the Monte Carlo search technique "Boltzmann Annealing" known as Adaptive Simulated Annealing (or Very Fast Simulated Re-Annealing). Simulated Annealing code is generally available to those skilled in the art over the Internet at http://www.ingber.com. The code provides operational steps for rapidly searching a large parameter space for an optimal solution given a cost function. An explanation of Simulated Annealing is found in Ingber, A.L. "Very Fast Simulated Re-Annealing" J. Math. Comput. Modeling (1989) 12: 967-973.

- 12 -

The simulated annealing technique employed in the invention uses algorithms which are well-known to those skilled in the mathematical arts in the following novel fashion. Coefficients of stretch and shift are at first randomly selected and applied to the normalized data traces. The cost is calculated. New coefficients are then selected within a range defined

5   by the annealing schedule (or temperature function "T") which governs the amount by which coefficients may be changed with each trial. The new coefficients are applied and cost is again determined. If the value of cost is lower than before, then the new point is used as the starting point for the next calculation. If the value of cost is higher, then the original coefficients are usually used again as the starting point of the next calculation. As in all simulated annealing

10   processes, however, there is a finite probability P (initially on the order of about 20% or less) that the higher cost value will be used as the starting point. As the number of calculated values increase, T and P are reduced, thus tending to localize the search space around an area of low cost. Annealing temperature schedules allow the "temperature" parameter to be raised to a higher value again at intervals during the search, emulating the process of annealing used

15   to heat treat metals. Eventually, when the search is fully completed, simulated annealing theory argues that the lowest cost value parameters will be found. In the above method, approximately 5000 sets of parameters are tested per calculation, representing 0.00001% of the available parameter field

In practice, the "fast annealing" modification of Ingber is found to be satisfactory to

20   obtain cost values low enough to base-call data from the Visible Genetics MicroGene Blaster™ (see Ingber, A.L. "Very Fast Simulated Re-Annealing" J. Mathl. Comput Modeling (1989) 12:967-973).

After the successful determination of the best fit parameter set for a given window of data points, the next window of data is analyzed. The next window is selected to be the same

25   number of data points as the first window, with an overlap of about 50% with the first window. Again, the coefficients of shift and stretch are identified which provide the lowest cost value when applied to the signal output functions. Thus, coefficients are determined for a series of piecewise domains, e.g., piecewise linear or cubic domains. The process of selecting windows and calculating coefficients continues until all the data has been analyzed.

30   When calculating the coefficients of stretch and shift for windows after the first window, it is sometimes advantageous to use the coefficients from a neighboring window as

- 13 -

the starting point for the simulated annealing process since the coefficients for neighboring windows tend to be related. In such cases, the annealing schedule T can be much shorter, for example testing only 1500 sets of coefficients as opposed to 5000. In addition, because it is desirable that the warp functions generated be continuous, subsequent windows may in fact be evaluated as two "sub-windows." In the first sub-window, stretch coefficients are constrained such that the warp function does not change the offset already established at the center of the previous full window. In the second sub-window, the stretch is allowed to vary in a narrow range.

Ultimately, by this process of sliding a window in overlapping steps across the combined normalized data, a "warp function" is arrived at for each normalized data trace. This function reflects the relationship between optimum-cost shift values for each window. By connecting the determined points of the function, a curve is defined which gives coefficients of shift for each point in a data trace, and reflects stretch at each location by the slope of the curve.

As a general matter, warp functions are generated in this way for each data trace. Thus, if the coefficients are determined with reference to a fixed standard, four different warp functions reflecting the varying ccefficients needed to align the four data traces are generated. In practice, however, it will be understood that the warp functions may be determined relative to one of the four data traces. In this case, the coefficients for the one fixed data trace will all be one. In the context of this application, the phrase "generating a warp function for each normalized data trace" encompasses both of these embodiments. The warp function may be represented by a plot of alignment of data points of the three warped traces, e.g., C, G and T against A as shown in Fig. 8A or as a plot of displacement from the A trace versus data point as shown in Fig. 8B.

Each warp function is applied to its respective raw or conditioned data trace to adjust the alignment of the data trace and generate a "warped data trace." The four warped data traces are then combined in alignment to produce an aligned data set.

An additional peak spacing warp function may be generated and applied to adjust for variations in peak spacing as part of creating the warped data traces. The peak spacing warp function, and also the base separation function useable as a standard trace, are generated by

- 14 -

testing postulated peak separation values in successive windows of the data trace and minimizing a peak separation cost function for each window. A suitable cost function is

$$COST = \sum d^2$$

where d is the distance between each actual peak and a hypothetical peak located at the position fixed by the postulated peak separation value.

5   Presentation of the aligned data set may be done internally within a computer for use with base-calling functions, or it may involve display of the aligned data set on a video monitor. Either way, the presentation allows further use to be made of the modified output signals, for base-calling and other purposes. For example, the video display of aligned data may be useful to permit an operator to make manual adjustments, and to observe inaccuracies

10   in base-calling.

Base-calling on the aligned data set may be performed in a variety of ways, including those base-calling techniques described in Tibbetts and Bowling (US Pat. No. 5,365,455) and Dam et al (US Pat. No. 5,119,316). Two preferred approaches to base-calling are described in detail below.

15   In the first approach for base-calling, peaks in each warped data trace making up the aligned data set are identified in the same manner in which peak detection was performed prior to normalization of the data traces. A minimum peak height from the base-line may be selected by the operator to avoid spurious results. Identified peaks are then used for base-calling.

20   Occasionally, peaks may represent a plurality of bands. It is necessary to determine which peaks these are, and how many bands they represent. An excellent method to employ is gaussian deconvolution whereby a peak is deconvolved into one or more standard gaussian peaks representing singleton peaks. It is found that peaks generated from DNA sequencing reactions using T7 polymerase (Pharmacia, Sweden) and Thermo Sequenase™ (Amersham

25   Life Sciences) generate the most consistent gaussian peaks.

The standard gaussian peak is determined as shown in Fig. 9. Peaks are located in a conditioned data trace from one channel. A line is drawn between peak points. The point on the line halfway between peaks is joined to the data trace by a line L perpendicular to the baseline. The area under the curve A and between the two perpendicular lines (L(x), L(x+1))

- 15 -

is determined. Height (h) is measured from the baseline to the peak. h and A are used to calculate sigma (σ) according to the equation:

$$\sigma = \frac{A}{h\sqrt{\pi}}$$

where σ represents the distance on the x-axis between the peak and the point at which the
5   value of the gaussian function

$$y = e^{-(x/\sigma)^2}$$

equals $1/e$.

For each detected peak, σ is determined. For those peaks where the length of both L(x) and L(x+1) are greater than ½ h, a linear regression is performed on the value of σ.
10   Statistically, at most about 25% of peaks are expected to represent doubletons, triples or greater, so using a second linear regression to correct for a width trend over each window, the narrowest 50% of the peaks are selected for use in constructing a piecewise cubic "singleton width discriminant function" that specifies a model width of a singleton at each location in the data array. All peaks in the window that are narrower or equal to this function are deemed to
15   be singletons to a first approximation. This approximation may be further refined using constraints such as peak area, etc. (Fig. 10) In this way, a standard or model singleton gaussian peak height and width at any point on the data trace may be defined.

The characteristics of the standard gaussian peak(s) and the positions of the singletons found via the discriminant function in conjunction with the base separation function are then
20   used to classify all the peaks in the aligned data traces. The features (e.g., height, width and/or area) of the standard peak are compared to the features of a detected signal peak to determine the number of bases represented.

For example. if the standard spacing indicated by the base separation function is consistent with there being three peaks in a region between a pair of singletons and that region
25   is occupied by a large peak, the characteristics of the standard gaussian peak (area, height etc.)

- 16 -

are used to determine whether two, three or four peaks are most likely to be the number of base pairs represented by the large peak. Thus, for example, the difference between the area of the large peak and the area of the standard gaussian peak can be evaluated. If the area of the large peak is approximately three times the area of the standard peak, then the large peak is treated as representing three bases. Similarly, if the area is closer to four times the area of the standard peak, the large peak is treated as representing four bases. A similar, although less sensitive analysis can be performed based on the height of the peaks. The preferred analysis takes multiple peak features into account.

The primer peak may be eliminated or ignored as described for preparing the normalized signal output, so as not to interfere with base-calling.

Another method of identifying and eliminating the primer peak uses a peak counting method. The data stream is divided into windows of a certain number of data points. The peaks in each window are counted. When a primer peak is in the window, a window that normally would include 10 peaks, may have only 2 peaks. This window is eliminated from consideration, and other windows are used for alignment and base-calling.

Once the individual peaks are identified and the multiple peak curves are divided into individual gaussian peaks, the data may be base-called. Each peak is identified with one channel, representing a single base. Peaks are therefore assigned to specific bases, in sequential order, until the full sequence is identified.

Fig. 11 outlines a second approach to base-calling using the aligned data set of the invention. In this case, peaks are identified within the aligned data set, for example as discussed above, and the data set is then divided into segments or windows. Within each window, a subset of the peaks are selected, for example 45% of the peaks, using selection criteria which select those peaks which are most likely to be "singleton peaks," i.e., peaks which represent only a single base and where that base is not the same as either adjacent base such that there should be no compression with adjacent peaks.

The selection of singleton peaks follows a several step process in which the features of each peak, i.e., width, height and area etc. are extracted and plotted in n-space. Based on the distribution of the peaks and the fact that, statistically, 56% of the bases in any sequence should be singletons within this definition, the region of n-space most likely to contain singleton peaks is defined. The peaks which fall within this region of space are then collected.

- 17 -

From the first pass singleton classification, piecewise singleton height and width functions that estimate model singleton features are defined for each segment or window. These functions are then used to reclassify the peaks from the aligned data set as singleton or non-singleton peaks. From this refined singleton classification, a single piecewise single base

5   separation distance function that estimates base-to-base separation is determined and used to provide postulated positions for all bases. These positions, combined with the actual peak features, neighboring characteristics and raw signal characteristics are then used to provide a best estimate of base call, together with a measure of the certainty of that call at each postulated base positions.

10   The method of the invention is advantageously practiced using a dedicated apparatus for determining nucleic acid sequences. As depicted in Fig. 12. such an apparatus comprises a sequencer 90 having an electrophoresis gel holder 901 disposed between electrodes 902 which are used to apply an electric field to a gel placed in the holder to cause oligonucleotide fragments to migrate within the gel; a detection system comprising a source 908 for an

15   interrogating beam 92 and a detector 907 for detecting the passage of oligonucleotide fragments through a detection zone. for example by monitoring emitted light 99; and a data processing system 96 operatively connected to the detector 95 for receiving raw data traces for each of the four chain termination product mixtures for a sample. Suitable gel holders, electrodes and detection systems are disclosed in US Patent Application No. 08/353,932, PCT

20   Patent Application No. PCT/US95/15951, and US Patent Application No. US Patent Application No. 08/387,272, all of which are incorporated herein by reference. although it will be understood that the particular configuration of the electrophoresis and detection system is not critical to the present invention.

The data processing system is suitably a personal or mini-computer which has stored

25   therein a programmed instruction set effective to

identify peaks in each of the four data traces;

normalize the height of said peaks in each of said data traces to a common value to generate four normalized data traces;

combine the four normalized data traces in an initial alignment;

30   determine coefficients of shift and stretch for selected data points within each normalized data trace, said coefficients optimizing a cost function which reflects the extent of

overlap of peaks in combined normalized data traces to which the coefficients have been applied, said cost function being optimized when the extent of overlap is at a minimum;

generate a warp function for each normalized data trace from the coefficients of shift and stretch determined for the selected data points;

5   apply each warp functions to the respective data trace or the normalized data trace to produce four warped data traces; and

align the four warped data traces to form an aligned data set. The data processing system may be connected to a video display 97 for displaying the aligned data set.

In a preferred embodiment of the invention, the apparatus of the invention reports
10   confidence levels to the system operator for some or all of the bases identified in the sequence. The confidence level advantageously reflects both (1) the arithmetic agreement between the signal and the model, and (2) other features of the data signal (for example expanded peak width) which may indicate reasons that the confidence level should be lower than the apparent level based on arithmetic agreements. This confidence level can be reported for all peaks, or it
15   can be reported only for those peaks for which the confidence level falls below a selected threshold value. Peaks may also be flagged during the reporting process to report ambiguities in the identification of the number of bases represented by a multiple peak feature.

CLAIMS

1.   A method for aligning data traces from four channels of an automated electrophoresis detection apparatus, each channel detecting the products of one of four chain-termination DNA sequencing reactions, whereby said four channels together provide information concerning the position of all four bases within a nucleic acid polymer being analyzed, comprising the steps of:

(a)   identifying peaks in each of the four data traces;

(b)   combining the four data traces in an initial alignment;

(c)   determining coefficients of shift and stretch for selected data points within each data trace, said coefficients optimizing a cost function which reflects the suitability of a trial alignment resulting from the application of the trial stretch and shift coefficients to the combined data traces;

(d)   generating a warp function for each data trace from the coefficients of shift and stretch determined for the selected data points;

(f)   applying each warp functions to the respective data trace to produce four warped data traces; and

(g)   assembling the four warped data traces to form an aligned data set.

2.   The method of claim 1, further comprising the step of normalizing the data traces to generate four normalized data traces in which homozygous peaks are all of substantially equal height prior to combining the four data traces into an initial alignment.

3.   The method of claim 1 or 2, further comprising the step of displaying the aligned data set on a video display.

4.   The method of any of claims 1 to 3, wherein the cost function reflects the extent of overlap of peaks in the combined normalized data traces to which the coefficients have been applied.

1   5.   The method of any of claims 1 to 4, wherein the cost function

2   determines the area of a region above the combined normalized data traces and below the

3   common value, and wherein the cost function is optimized when this area is at a minimum.

1   6.   The method of any of claims 1 to 4, wherein the cost function

2   determines the area of a region below the combined normalized data traces, and wherein the

3   cost function is optimized when this area is at a maximum.

1   7.   The method of any of claims 1 to 3, wherein the cost function

2   determines the area of a first region above the combined normalized data traces and the area

3   of a second region, said second region being below the highest-edge of the combined

4   normalized data traces and above the second highest edge of the combined data traces,

5   wherein the cost function is optimized when the first area is minimized and the second area is

6   maximized.

1   8.   The method of any of claims 1 to 7, wherein the coefficients of shift and

2   stretch for selected data points within each normalized data trace that yield the optimum value

3   of the cost function are determined using a process of simulated annealing.

1   9.   The method of any of claims 1 to 8, further comprising the step of

2   conditioning the data traces prior to identification of the peaks in the data traces.

1   10.   A method for determining the sequence of a nucleic acid polymer

2   comprising the steps of:

3   (a)   obtaining data traces from four channels of an automated electrophoresis detection

4   apparatus, each channel detecting the products of one of four chain-termination DNA

5   sequencing reactions, whereby said four channels together provide information

6   concerning the position of all four bases within a nucleic acid polymer being analyzed;

7   (b)   aligning the data traces by a method according to any of claims 1 to 9 to produce an

8   aligned data set;

9   (c)   evaluating the aligned data set to determine the sequence of bases within the nucleic
10   acid polymer.

1   11.   The method according to claim 10, further comprising the steps of

2   determining standard gaussian peak shapes for data points along the aligned

3   data set;

4   assigning peaks in the aligned data set as singleton peaks or multiple peaks by

5   comparison of the peaks in the aligned data set to the standard gaussian peak characteristics;

6   and

7   determining how many standard gaussian peaks are contained within each

8   multiple peak, wherein each multiple peak is treated as that number of singleton peaks for

9   purposes of base-calling.

1   12.   The method according to claim 10 or 11, further comprising the step of

2   reporting confidence levels for at least of portion of the called base peaks.

1   13.   The method according to claim 12, wherein the confidence level reflects

2   the arithmetic agreement between the peak in the data trace and a model, and those features of

3   the peak which might justify assignment of a lower confidence level.

1   14.   An apparatus for determining the sequence of a nucleic acid polymer

2   comprising:

3   (a)   an electrophoresis gel holder;

4   (b)   first and second electrodes disposed to apply an electric field to the

5   electrophoresis gel disposed within the electrophoresis gel holder to cause oligonucleotide

6   fragments loaded on the electrophoresis gel to migrate within the electrophoresis gel;

7   (c)   a detection system comprising an interrogating beam and a detector for

8   detecting the passage of oligonucleotide fragments through a detection zone; and

9   (d)   a data processing system operatively connected to the detector for

10 receiving four data traces, one for each of four chain termination product mixtures for the

11 nucleic acid polymer, wherein the data processing system has stored therein a programmed
12 instruction set effective to

13   identify peaks in each of the four data traces;

14   combine the four data traces in an initial alignment;

15   determine coefficients of shift and stretch for selected data points within each
16 data trace, said coefficients optimizing a cost function which reflects the suitability of a trial
17 alignment resulting from the application of the trial stretch and shift coefficients to the
18 combined data traces;

19   generate a warp function for each data trace from the coefficients of shift and
20 stretch determined for the selected data points;

21   apply each warp functions to the respective data trace to produce four warped
22 data traces; and

23   assemble the four warped data traces to form an aligned data set.

1   15.   The apparatus according to claim 14, further comprising a video display
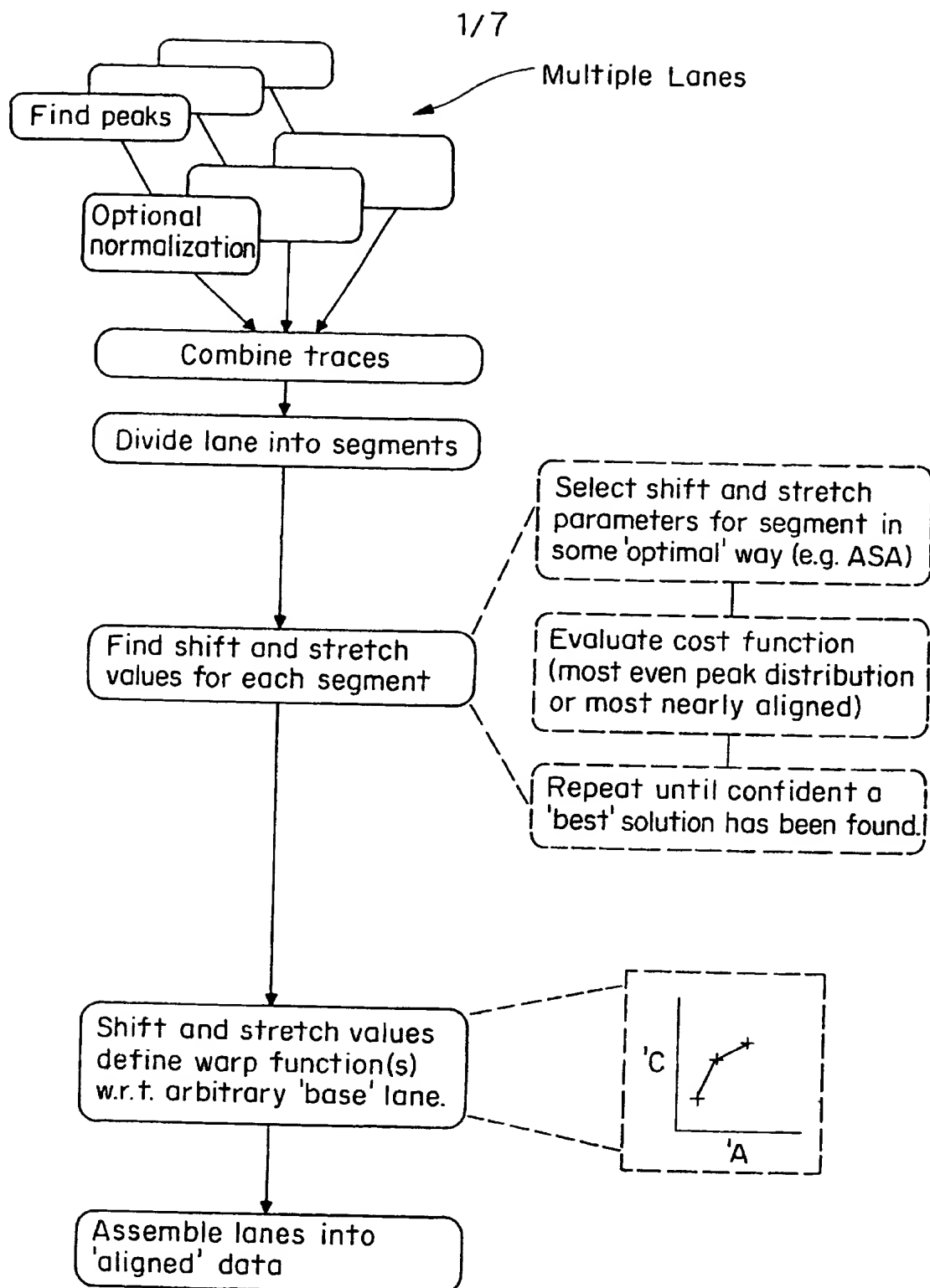2   for displaying the aligned data set.

Multiple Lanes

Find peaks

Optional
normalization

Combine traces

Divide lane into segments

Find shift and stretch
values for each segment

Select shift and stretch
parameters for segment in
some 'optimal' way (e.g. ASA)

Evaluate cost function
(most even peak distribution
or most nearly aligned)

Repeat until confident a
'best' solution has been found.

Shift and stretch values
define warp function(s)
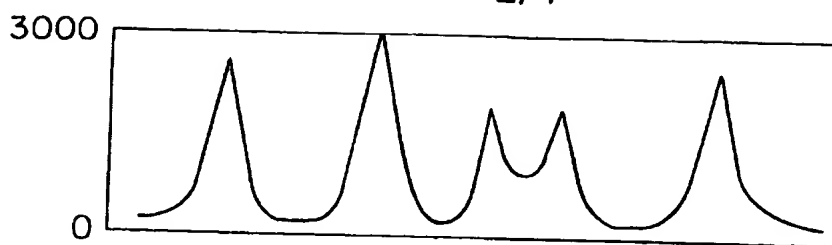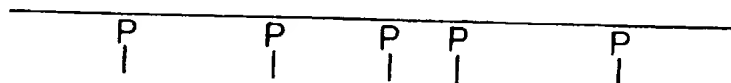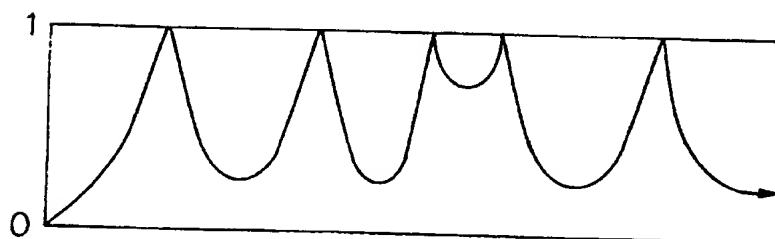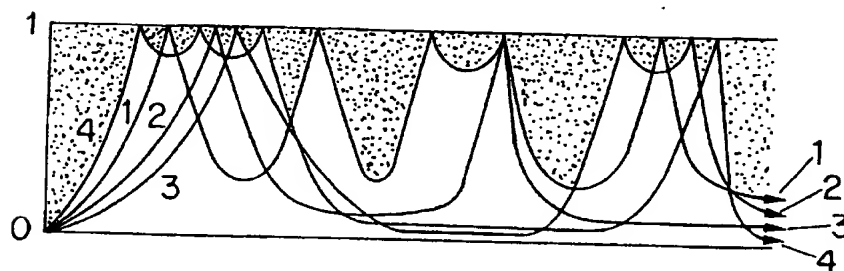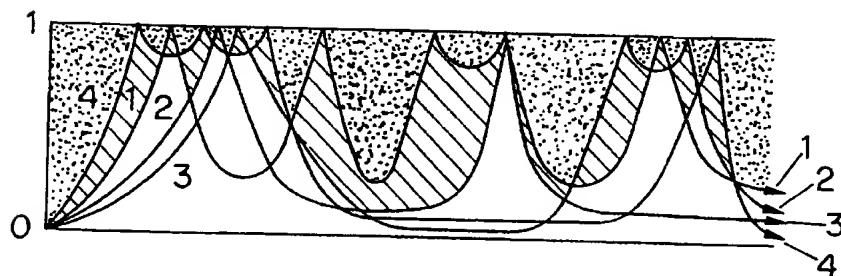w.r.t. arbitrary 'base' lane.

'C

'A

Assemble lanes into
'aligned' data

FIG. 1

FIG. 2

FIG. 3

FIG. 4

FIG. 6A

FIG. 6B

FIG. 7

FIG. 5A

FIG. 5B

C, G, T
DATA POINT

NO SHIFT OR STRETCH

A (DATA POINT)

FIG. 8A

DISPLACEMENT FROM A LANE
IN DATA POINTS

☐ T
○ G
△ C

DATA POINT NUMBER

FIG. 8B

FIG. 9

TIME ⟶

CONDITIONED DATA TRACE

FIG. 10

Four lanes aligned for most even peak distribution (from alignment algorithm).

↓

Extract peaks

↓

Divide lane into segments

↓

Select the 45% of peaks in each segment most likely to be singleton peaks (statistically 75% of peaks in each lane should be singletons).

↓

From the first pass singleton classification define piecewise singleton height, and width functions that estimate model singleton features at each sample location in the lane.

↓

Repeat the singleton classification using descriminants centered on the singleton feature functions.

↓

From the refined singleton classification define a piecewise single base separation distance function that estimates base to base separation at each sample location in the lane.

↓

Use model singleton peak spacing function to provide postulated positions for all bases.

↓

Use model singleton feature (width, height, etc) functions, peak features, neighbouring peaks and raw signal characteristics to provide best estimate of base call, plus the certainty of that call, at each postulated base position.

Extract peak features (width, height, area, etc.)

↓

Plot peak features in 'n-space'

↓

Define the region of n-space most likely to contain singleton peaks (e.g. narrowest, shortest, etc while discarding outliers)

↓

Collect the peaks whose features fall in the defined area

For all reasonable base separations, sum the square of the distance from each singleton peak to the model location. Choose the separation distance for the segment as the one with the least squared distance

FIG. 11

FIG. 12

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC 6   G01N27/447

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6   G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GB 2 225 139 A (UNITED KINGDOM ATOMIC ENERGY AUTHORITY) 23 May 1990 see abstract; figure 3	1,10,14
A	US 5 419 825 A (H. FUJII) 30 May 1995 see abstract; figure 1	1
A	EP 0 198 403 A (FUJI PHOTO FILM, CO, LTD.) 22 October 1986 see abstract; figure 1	1
A	EP 0 476 712 A (FUJI PHOTO FILM CO., LTD.) 25 March 1992 see abstract; figure 4	1
A	US 5 273 632 A (T. G. STOCKHAM) 28 December 1993 see abstract; figure 1	1

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 September 1997	-2. 10. 97

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl, Fax (+ 31-70) 340-3016	Duchatellier, M

INTERNATIONAL SEARCH REPORT

Information on patent family members

Interna | Application No

PCT/CA 97/00463

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2225139 A	23-05-90	NONE	
US 5419825 A	30-05-95	JP 2105211 C	06-11-96
		JP 5107227 A	27-04-93
		JP 8027258 B	21-03-96
EP 198403 A	22-10-86	JP 1776498 C	28-07-93
		JP 4067629 B	28-10-92
		JP 61233369 A	17-10-86
		JP 1776499 C	28-07-93
		JP 4067630 B	28-10-92
		JP 61233370 A	17-10-86
EP 476712 A	25-03-92	JP 1933534 C	26-05-95
		JP 6064058 B	22-08-94
		JP 63191959 A	09-08-88
		JP 1816323 C	18-01-94
		JP 5029070 B	28-04-93
		JP 62225956 A	03-10-87
		DE 3751366 D	27-07-95
		DE 3751366 T	02-11-95
		DE 3781307 A	01-10-92
		EP 0242629 A	28-10-87
		US 4885696 A	05-12-89
US 5273632 A	28-12-93	NONE	